DOCUMENT RESUME

ED 380 483 TM 022 802

AUTHOR Jones, Russell W.
TITLE Performance and Alternative Assessment Techniques: Meeting the Challenge of Alternative Evaluation Strategies.
PUB DATE Jul 94
NOTE 29p.; Paper presented at the International Conference on Educational Evaluation and Assessment (2nd, Pretoria, Republic of South Africa, July 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Assessment; Educational Trends; Evaluation Methods; Knowledge Level; *Multiple Choice Tests; *Portfolio Assessment; Portfolios (Background Materials); *Test Construction; Test Format; *Testing
IDENTIFIERS *Alternative Assessment; *Performance Based Evaluation

ABSTRACT

One of the most influential contemporary trends in educational evaluation in the United States is the move away from traditional testing methods toward "authentic assessments," which are designed to measure student performance of skills, abilities, and knowledge directly. While there is no consensus as to precisely what constitutes authentic assessment, there is general agreement that it incorporates: (1) emphasis on examinee performance, assessing not only what the examinee knows, but what the examinee can do; (2) use of direct methods of assessment; (3) inclusion of a high degree of realism; and (4) activities for which there may be no one correct answer, in a simulation of realism. The primary distinction between traditional testing methods and authentic assessment is the choice of question format. Certain segments of the educational community have called for a move from the multiple-choice format to question formats considered to reflect higher-order cognitive processes more accurately. This has resulted in the development or adoption of a broad range of formats, including standardized patient, audio-visual context setting, computer-based problem solving, multiple choice with justification, latent image, performance, and portfolio. Two figures illustrate the discussion. (Contains 30 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RUSSELL W. JONES

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Performance and Alternative Assessment Techniques:

Meeting the Challenge of Alternative Evaluation Strategies

by

Russell W. Jones

Center for the Study of Testing, Evaluation and Educational Policy
Campion Hall, Boston College, Chestnut Hill, MA 02167, U.S.A.
(Internet: jonesru@hermes.bc.edu)

ED 380 483

July 1994

Keynote address to the 2nd International Conference on Educational Evaluation and Assessment. Pretoria, Republic of South Africa.

TM022802

Performance and Alternative Assessment Techniques: Meeting the Challenge of Alternative Evaluation Strategies

by

Russell W. Jones
Boston College

Abstract

One of the most influential contemporary trends in educational evaluation in the United States is the move away from traditional testing methods towards "authentic assessments." These assessments, which take a variety of forms, have been called "authentic" because they are designed to measure directly student performance of skills, abilities and knowledge that are valued in and of themselves (i.e., they measure authentic performance). Although, beyond this broad philosophical approach, there exists no consensus as to precisely what constitutes authentic assessment, there is general agreement that authentic assessment incorporates the following four factors (Horvarth, 1991; Jones & Hambleton, 1992; Royer, Carlo, & Jones, 1993): (1) the assessment emphasizes examinee performance, thereby assessing not only what an examinee knows, but also what the examinee can do; (2) the assessment uses direct methods of assessment; (3) assessments incorporate a high degree of realism; and (4) to reflect realistic situations assessments may include activities for which there exists no single correct answer, may consist of group assessment rather than individual assessment, and may be continuous over time.

The primary distinction between what are considered traditional testing methods and authentic assessments is the choice of question format. Hence, the recent move towards authentic assessment has resulted in a great deal of attention being focused on question formats. In particular, certain segments of the educational community have called for a move away from the multiple-choice format, which has dominated much of testing in the United States (and, according to Kellaghan and Greaney (1992), multiple-choice questions are increasingly being used in this, the African continent), to question formats considered

to more directly measure performance and which may be considered to more accurately reflect higher order cognitive processes. This has required the development or adoption of a broad range of question formats including standardized patient, audio-visual context setting, computer-based problem solving, multiple-choice with justification, latent image, performance and portfolio. The development and implementation of these question formats holds the promise that soon it will be possible to readily assess higher order cognitive skills and abilities such as reasoning, problem solving and critical thinking. Assessment of these skills and abilities is a potential source of a wealth of information for both educational and psychological purposes. Moreover, these are the cognitive outcomes in which many educators profess to be most interested (Jones & Hambleton, 1992).

Performance and Alternative Assessment Techniques:

Meeting the Challenge of Alternative Evaluation Strategies

One of the most influential contemporary trends in educational evaluation within the United States is the move away from traditional testing methods towards "authentic assessments." These assessments, which take a variety of forms, have been called "authentic" because they are designed to measure directly student performance of skills, abilities and knowledge that are valued in and of themselves (that is, they measure authentic performance). Conversely, traditional testing relies on indirect measures of performance. The primary distinction between what are considered authentic assessments and traditional testing methods is the choice of question format. Hence, the recent move towards authentic assessments has resulted in a great deal of attention being focused on question formats. The cause of this movement away from traditional testing and towards authentic assessment is a general and widespread dissatisfaction with the current educational performance of schools within the United States. Although, this dissatisfaction is directed towards most aspects of school education including school curricula, organization and teacher training, evaluation is the theme of this conference and evaluation will be the subject of this address.

Dissatisfaction with Standardized Tests

A cursory review of any classroom curriculum or school district assessment plan will reveal testing to be an extremely common educational practice. A recent report by the National Commission on Testing and Public Policy estimated students in primary and secondary schools within the United States take 127 million separate standardized tests annually at a cost of approximately 1 billion dollars (National Commission on Testing and Public Policy, 1990). This quantity may seem enormous yet there is a continued call for even more testing as the use of test results expands to include educational accountability of

not only students, but also the accountability of teachers, schools, educational programs, school districts and even states (see, for example, the State of Mathematics Achievement: NAEP's 1990 Assessment of the Nation and the Trial Assessment of the States, Mullis, Dossey, Owen & Phillips, 1991) and entire countries (see, for example, the national testing program, the National Assessment of Educational Progress, or the Third International Math and Science Study currently comparing 50 countries, or the first or second International Assessment of Educational Progress). As greater use is being made of test results in educational accountability, so there has been a concomitant increase in the degree of interest displayed by those who are affected by assessment results. Most especially people and organizations who are affected by these tests want to ensure the tests themselves are fair and accurately measure what students are learning (Hambleton, Jones & Cadman, 1993).

Question formats may be divided into those that are objective and those that are subjective. Objective formats are by far the most prevalent, at least within the United States, and dominate traditional standardized testing programs. They include multiple-choice, true-false, matching and multiple true-false questions. These are called objective formats because such questions can be marked objectively, that is, the same mark will be allocated for an examinee response regardless of who marks the response. Conversely, subjective question formats require the subjective judgment of the examiner to enter into the evaluation process. Hence, the mark for any examinee response may differ from one examiner to another, and, under certain circumstances, the same examiner may even differentially mark two identical responses (Gronlund, 1982). Within the United States the vast majority of traditional tests have used objective question formats, most especially the multiple-choice question format. (Also, according to a recent study by Kellaghan and Greaney [1992], multiple-choice formats are increasingly being used in this, the African continent.)

Criticisms of Traditional Objective Tests. A review of the literature pertaining to authentic assessment has uncovered 10 criticisms of objective tests (Hambleton, Jones, & Cadman, 1993). To wit, objective tests are said to:

- Foster a one-correct answer mentally.
- Narrow the curriculum.
- Focus on discrete skills.
- Under-represent the performance of lower socio-economic examinees.
- Are unable to measure higher-order thinking skills.
- Focus on product at the expense of process.
- Require students to select answers rather than create them.
- Lack construct validity (for example, tests are timed and they assess individual rather than group effort).
- Have negative effects on students and teachers.
- Encourage "teaching to the test".

Regardless of whether or not these criticisms are valid, they have added impetus to the movement away from traditional objective question formats used by standardized testing programs in the United States and Canada. Additional impetus has come from the growing realization that traditional standardized tests using objective question formats are not capable of measuring all of what schools expect students to learn. Nor are these testing formats capable of comprehensively evaluating all important higher-order cognitive outcomes (Hambleton, Jones & Cadman, 1993). Hence, those standardized testing programs employing objective question formats, such as the multiple-choice question, are coming under increasing pressure to develop or adopt alternative assessment techniques. These alternative assessments require students to use problem solving, reasoning and other higher order cognitive skills to engage in complex assessment tasks, rather than to merely

regurgitate discrete knowledge and exhibit an understanding of when it is appropriate to apply that knowledge. Specifically, there is a call for authentic assessment.

What is Authentic Assessment?

As mentioned earlier authentic assessments are so called because they are designed to measure directly student performance of skills, abilities and knowledge that are valued in and of themselves. Yet, beyond this broad philosophical approach, there exists no consensus as to precisely what constitutes authentic assessment. The controversy surrounding authentic assessment has led some educators to state, in recent times, the belief that almost any question not using the multiple-choice paper-and-pencil format can be considered an authentic assessment. Other writers have argued even carefully written, judiciously applied multiple-choice questions can qualify as authentic assessment (Royer, Carlo, & Jones, 1994). There is, however, general agreement that authentic assessment incorporates the following four factors (Horvarth, 1991; Jones & Hambleton, 1992; Royer, Carlo, & Jones, 1994): (1) the assessment emphasizes examinee performance, thereby assessing not only what an examinee knows, but also what the examinee can do; (2) the assessment uses direct methods of assessment; (3) assessments incorporate a high degree of realism; and (4) to reflect realistic situations authentic assessments may include activities for which there exists no single correct answer, may consist of group assessment rather than individual assessment, and may be continuous over time.

The movement away from traditional testing and the search for alternative assessment techniques has also led to a change in the goals of assessment (Royer, Carlo, & Jones, 1994). According to Arter and Spandel (1992, p. 36) to meet these new goals assessments should:

- Capture a richer array of what students know and can do than is possible with multiple-choice tests.

- Portray the process by which students produce work.
- ...align with what we consider important outcomes for students in order to communicate the right message to students and others about what we value.
- Have realistic context for the production of work, so that we can examine what students know and can do in real-life situations.
- ...chronicle development, give effective feedback to students, and encourage students to observe their own growth.
- ...improve achievement rather than just monitor it.

Clearly then, the search for alternative assessment techniques is an important and far reaching contemporary trend in educational evaluation and assessment. The remainder of this address will focus on the emerging alternative assessment formats that are thought, by many educators, to better evaluate authentic performance.

"Authentic" Question Formats

Several question formats lend themselves to evaluate examinees through direct evaluation of examinee performance. These include standardized patient, audio-visual context setting, computer-based problem solving, multiple-choice with justification, latent image, performance and portfolio.

Standardized Patient. Initially designed for use within the medical community this format has now been adopted by many professions involving clinical situations (for example, dentistry, nursing and psychology). An actor or actress is trained to display specific symptoms of a particular condition, i.e., to become a "standardized patient." The "patient" is then introduced to a candidate who must examine and question the "patient" to determine the nature of the problem and prescribe appropriate treatment. One or more examiners observe the behavior of the candidate and mark the performance of the candidate. This is a

particularly useful format because during the course of the administration of a single question a number of different aspects of the practitioner-client relationship can be evaluated. This makes full use of the resources, time and effort required in the assessment situation.

The standardized patient format is particularly suitable for assessment of those tasks requiring contact with others, such as those commonly occurring in medicine, nursing, clinical and counseling psychology, and dentistry. However, this format may also be used within education. Standardized patient questions are especially useful when evaluating certain members of school staff. For example, the ability of a Career Guidance Officer or Counselor to develop a program of study to match a student's desired career path may be assessed by an actor or actress taking on the role of student whilst asking specific questions and demonstrating specific behavior. The candidate Career Guidance Officer or Counselor is required to respond appropriately to these questions and behavior. Similar techniques may be used to assess school psychologists, school medical staff and school counselors. The standardized patient format is not limited to evaluating school support staff. A teacher's response to an actor or actress portraying the role of a student who exhibits specific behavior may also be assessed using the standardized patient format. Students too, can be assessed using this question format. The mock interview, often staged during career counseling classes, is an ideal situation in which to incorporate standardized patient questions to evaluate student behavior. In this case the actor or actress becomes the potential employer who interviews a potential employee (the student). In each of these examples one or more examiners observer and mark candidate performance. The reality of the performance can be enhanced by administering the assessment in the location where the "real" event would occur and by using realistic props.

Subjectivity of the marking procedure within standardized patient questions, and indeed within all subjectively marked question formats, can be reduced by requiring the

examiners to mark the candidate on a standardized scoring form. This form should contain details of pre-determined criteria which the examinee is required to meet and which have been established earlier by a panel of suitably qualified and experienced personnel. Furthermore, a panel of several examiners should be used to mark responses rather than a single examiner. This increases the objectivity of the mark and permits continual monitoring of reliability.

The advantages of this format are many and include the concurrent assessment of more than one skill, or of complex composite skills, and the very practical essence of the question which closely reflects events the candidate will likely meet in practice/real life. However, thorough training is required of the actor or actress in all aspects of their performance. Moreover, careful control must be exerted over the standardized patient to ensure they always present a consistent performance to each candidate and effects such as fatigue or boredom or a like or dislike of particular candidates do not bias the examination one way or another. Vu and Burrows (1994) also highlight the high cost of developing, administering and marking assessments using standardized patients and the fact that these assessments yield only moderate reliability.

Audio-Visual Context Setting. This is a very realistic question format which creates the performance scenario through the use of audio and/or visual stimuli. Technological advances, such as the widespread availability of the video camera/recorder and VCR, have made it comparatively easy to film scenarios and present them to candidates in order to "set the stage" for an authentic examination. Indeed, the responsive ability of videodisk technology, which can provide the operator with a choice of options and immediate feedback regarding any option they choose, make it possible for examinees to be presented with an audio-visual representation of a scenario from which they are required to make a decision. Feedback regarding the outcome of this decision can be immediately provided to the candidate in the form of a modified scenario which continues to unfold until the

candidate is required to make another decision, whereupon immediate feedback is again provided. An examination can consist of one or several of these realistic scenarios through which the candidate is required to successfully work in order to provide a satisfactory performance.

As an example, let us consider a teacher certification test administered to potential science teachers. A great many teacher certification examinations have been rightly criticized for their inability to simulate the varied and complex environment encountered within the classroom. Through the use of a question incorporating audio-visual context setting the reality of the question can be greatly enhanced. A teacher applying for certification can view a television monitor where they are shown a classroom scenario recorded on a videodisk. Such a scenario could involve a science experiment using Bunsen burners whereby one student has placed the flame of the burner immediately beneath a shelf of books (see Figure 1). At certain intervals during the scenario the presentation pauses

---

Inset Figure 1 About Here

---

and the candidate is asked to select a desired course of action from a series of options in response to what they have observed unfold in the classroom situation. The candidate is given a series of choices from which to choose, such as (in this example) (1) stopping the entire class, (2) rushing over to intervene, (3) ignoring the situation, or (4) calling the attention of the student to the hazard. The candidate enters their desired choice of action (possibly by the use of a computer mouse or pen sensitive computer monitor) and the candidate is then shown the next sequence of events determined in part by the course of action selected by the candidate. If the candidate chose to ignore the situation the next phase in the scenario could show the book shelf beginning to smolder. Again, the scenario pauses and the candidate is once more required to select a course of action. An examiner can monitor the candidate's performance and then decide if the candidate has acted

appropriately and therefore demonstrated sufficient mastery of the skill(s) required by the relevant objective.

Audio-visual context setting offers the potential for increased face validity compared to traditional methods, and the possibility to evaluate complex and interacting cognitive skills. The audio-visual apparatus may also be programmed to evaluate the performance of the candidate, thereby increasing objectivity and avoiding the need for an expert examiner to be present for the purposes of marking (although, of course, a proctor would still be required). As an aside, this format presents an additional issue when marking candidate responses (Jones & Hambleton, 1992). This issue is whether to impose a penalty upon the candidate if their performance indicates hazardous behaviour. For example, how should the candidate for science teacher certification, in the previous paragraph, be marked if they fail to act appropriately and the school burns down? Clearly, the candidate may be expected to fail the question, but should consideration also be given to his/her failing the entire test?

Computer-Based Problem Solving. This format makes use of computers to present an examinee with information about a hypothetical problem and the examinee is tasked with finding a solution (Melnick, 1990). This format may be adapted to many assessment situations including those involving practitioner-client interaction similar to the standardized patient format. In this context these questions are called "computer-based clinical situations." Again, the examinee is required to diagnose a client's condition and prescribe treatment or an appropriate course of action. While computer-based clinical situations do not offer the same degree of reality as the standardized patient, they do offer the advantages of greater comparative economy and consistency when providing the candidate with information. Such is not always the case with trained actors who, in addition to being expensive and time-consuming to train, are also prone to fatigue and to personal likes or

dislikes of candidates which may interfere with the consistency of presentation during the course of the testing of consecutive candidates.

Of course, computers can be just as effectively used to provide hypothetical situations for candidates whose evaluations do not involve clinical situations. Realistic hypothetical scenarios can be presented to candidates who are required to be evaluated regarding their performance on a wide range of objectives from many different curricula. The candidates can then be evaluated on the basis of how well they perform the appropriate tasks as a consequence of the prompting they receive from the computer-generated scenarios. In addition, the reliance on hypothetical situations can be replaced by the re-enactment of historically real events. Computers can be used to economically recreate a particular event and a candidate asked to solve a problem which has a known, not hypothetical, outcome (see Figure 2). A graphic illustration is a pilot training exam whereby a computer within a

---

Inset Figure 2 About Here

---

flight simulator can present a pilot with the events leading to known airline disasters. This would allow the examiner to determine if the trainee was likely to make decisions or take actions which might also lead to disaster. Obviously, such a simulation has valuable instructional applications. The potential value for questions of this type as an aid to history assessment and instruction is also very promising. Another example is a geography or science curriculum concerning meteorology. Students can be exposed to real-life presentations of weather patterns and meteorological events whilst being required to forecast. The known outcome of these events permits immediate feedback to students regarding the accuracy of their forecast. Similarly, geology students can be shown speeded

computer simulations of geological events. In actuality these events may have taken many millennia whereas a computer can show simulations of these events in only seconds.[1, 2]

Multiple-Choice With Justifications. Not all authentic question formats place a heavy emphasis on the use of modern technology or simulation. One such example is the multiple-choice with justification format. This question format retains many of the benefits of the multiple-choice format (Benefits of the multiple-choice format include: their ability to assess a diverse range of curricula, abilities and psychological processes; the fact that many objectives can be measured quickly and easily using a series of multiple-choice questions; objective marking leading to high reliability; and, tests can be rapidly marked permitting prompt feedback to both examiners and examinees) but gathers additional information from the candidate by requiring the examinee to provide a brief written justification of his/her answer choice. This question format is particularly useful for formative assessments because invaluable information is obtained regarding incorrect reasoning, misconceptions, and gaps in the knowledge base of candidates.

---

[1] Computer-based simulations contain all the advantages inherent within any computer-based testing system. They include the potential for computerized adaptive testing; greater flexibility regarding the time and location of test administration (that is, it may no longer be necessary to bring examinees together at a specific time and place for the administration of a single paper-and-pencil test, but instead examinees may be able to sit a computerized exam at any computer terminal and at any time); the use of computerized question banks so questions may be drawn from a large pool of questions each with known psychometric properties; and, greater test security.

[2] Machine Marking. Many question formats are machine markable, that is, examinee responses are able to be evaluated by a machine. The "machine" may be as simple as an electronic scanner which can mark the responses made by an examinee on a standardized answer sheet or as complex as a built-in evaluation device written into a computer simulation program such as the program discussed under the "Computer Based Problem Solving" format. Marking by machine is valuable to the test developer because of the consistency, objectivity, and speed by which a machine can mark a large number of responses without the effects of confounding factors such as fatigue, boredom, or inconsistency between examiners. These effects can affect human markers with the subsequent introduction of error into the assessment process. Other advantages include the opportunity for rapid marking mechanisms capable of reporting the examination result immediately upon completion of the exam; greater reliability; and readily centralized and controlled test administration, recording and reporting of test marks (Bunderson, Inouye, & Olson, 1989).

Latent Image Question. This format is based on similar principles to that which will likely make the videodisk a valuable assessment tool in the near future. However, instead of a scenario being presented to the candidate via a series of related audio-visual presentations, the scenario is presented through a series of related written questions and background information. Each question is answered by the candidate who selects a specified number of options and, through the use of a special pen, each option reveals additional information to the candidate. Armed with this additional knowledge, the candidate moves on to the next question which is itself determined by the response to the previous question. Questions such as these are capable, to some extent, of mimicking real life situations where performance frequently relies on small packets of information provided via feedback, rather than the typical examination situation where an examinee is provided with all the information he/she requires to answer a question at the onset. Perhaps the greatest disadvantage of the latent image question format is the presentation of the scenario places a heavy reliance on the written word. Thus, examinee performance is, to a large extent, influenced by their reading ability.

Performance and Portfolio Question Formats

Of all question formats currently under consideration by members of the educational fraternity within the United States, the two receiving greatest attention are the performance and portfolio formats. Their comparatively more widespread adoption may be for several reasons including their intuitive appeal with powerful face validity, and because other authentic assessment formats tend to place a greater reliance on expensive technology.

Performance. Perhaps the most rapidly expanding question format within the United States is the performance format. This format holds the promise of addressing many of the criticisms leveled at traditional testing formats. Indeed, this question format more

than any other is seen by many educators, at least those outside the testing industry, as the panacea of evaluation. Performance questions require candidates to perform a practical task related to test objectives. For example, the ability of 13 year-old students to design and perform a scientific experiment may be evaluated by presenting the student with a problem and requiring the student to solve the problem by designing and performing a suitable experiment. Similarly, the reading ability of a grade 6 student may be evaluated by requiring the student to read aloud a poem or other written passage. Performance is evaluated by one or more examiners who observe and mark the performance.

Detailed examples may serve to demonstrate the richness of the information obtained by a performance question. The following two examples are mathematics questions taken from the 1991 International Assessment of Educational Progress[3] (Semple, 1992).

Sample Performance Question 1

Task Descriptor:

To make a 15 gram lump of modelling clay from a larger lump of clay, using a two-pan balance and two masses.

Apparatus:

- Two-pan balance
- 20 gram mass
- 50 gram mass
- Lump of modelling clay

Student Instructions:

Make a 15 g lump of clay using the materials provided and explain how you did it. (Simple instructions on how to use the balance were provided.)

[3] Other countries which have included a performance component within their national educational monitoring include England and Wales, with the Assessment of Performance Unit (APU), and Scotland, with the Assessment of Achievement Programme (AAP) (Semple, 1992).

Sample Performance Question 2

Task Descriptor:

To measure the capacity of two plastic containers using a measuring cup.

Apparatus:

- A large container filled with water
- 2 smaller containers labelled A and B.
- 500 ml measuring cup graduated in 25 ml units
- Containers A and B were marked with a black line at 375 ml and 275 ml respectively

Student Instructions:

Fill containers A and B up to the black lines from the large container. Measure the amount of water in each container in millilitres, using the measuring cup.

Observing a student performing these two tasks would give the observer information regarding several student abilities. These include the student's problem solving ability, ability to develop and implement a strategy, arithmetic skill, knowledge of mass, knowledge of volume, understanding and ability to read scales, and motor coordination. Furthermore, these tasks may be completed by using more than one strategy. Observation of the particular strategy selected by a student will give clues as to the cognitive strategies the student is applying to perform each task. Also, students can, and do, learn while performing these tasks.

There is a tendency for some educators to foster the belief performance questions are a new innovation, however, performance questions are most likely one of the oldest question formats. Anyone who has spent time in the armed forces is familiar with the common practice of evaluating a serviceman or woman's marksmanship by requiring them to shoot a rifle at a target. Their marksmanship is evaluated through their ability to accurately and consistently hit the target. This performance question has been used

ever since the rifle became a common weapon within the arsenal of an armed force. It is difficult to imagine the Egyptians during the Early Dynastic Period of 3100-2686 BC not having a similar test for marksmanship with a bow and arrow or spear. Another example of a performance assessment administered during the 11th Century BC is documented by Mehrens (1992).

Performance assessment is increasingly becoming incorporated into tests in many educational and credentialling situations through the addition of a practical component to traditional multiple-choice exams. These components are called performance questions and require candidates to perform a practical task related to the test objectives. The complexity of the task varies enormously. For example, the task may be as simple as correctly spelling a word during a spelling test or as complex as successfully teaching a class in the case of a teacher certification exam.

Portfolios. Portfolio, in common with many of the terms associated with authentic assessment, is a term with no clear definition. For the purpose of this address the definition provided by Arter and Spandel (1992), in their instructional module on how to use portfolios published in Educational Measurement: Issues and Practice, will be used. These authors define a student portfolio as:

> a purposeful collection of student work that tells the story of the student's efforts, progress, or achievement in (a) given area(s). This collection must include student participation in selection of portfolio content; the guidelines for selection; the criteria for judging merit; and evidence of student self reflection." (p. 36)

Of course this definition can be expanded beyond the realm of the student to encompass many skills. Again, as with other authentic assessments, the idea of portfolio assessment is not new. Anyone familiar with the field of traditional visual art (e.g., painting) knows an artist often presents a prospective client with a "portfolio" of their work. This portfolio has always contained samples of the artists work.

Modern technology offers the potential to move the portfolio beyond merely documenting a student's written or pictorial work. For example, a student's reading ability can readily be documented by making an audio cassette recording of the student reading a passage. Such a recording made at the beginning of a school year can be compared to a similar recording made at the end of the school year. This would provide clear and demonstrable evidence of the level of improvement made during the year. Similarly, a video cassette recorder can be used to make a permanent audio-visual record of student performance on a wide range of assessments. For example, the performance of students on the mathematics and science tasks discussed in the preceding section concerning performance assessment may be readily documented by filming students performing these tasks.

Video cassette recorders may also be used as an aid to documenting school accountability. Thus, the portfolio concept for students may be expanded beyond the assessment of individual student performance. Just as a portfolio can be constructed to monitor student performance so a similar portfolio may be used to document school accountability. In place of the written prose, essays, paintings, audio and video cassette recordings within a student's portfolio, a portfolio constructed for the purposes of school accountability could contain written records of school attendance and student test marks, pictorial records of improvements in school infrastructure (such as modifications to canteens, technology upgrades, expanded libraries and so forth) and audio or video cassette recordings of staff development tasks, school board meetings, parent-teacher interactions, and school activities which influence the wider community (such as students planting trees as part of a conservation program, visiting an elder hostel, or clearing up rubbish around the community).

Thus far, although there has been considerable discussion about portfolio assessment, assessment systems have been cautious about the wholesale adoption of portfolios. Three exceptions in the United States are the state-wide assessments in

Vermont, New Mexico and Kentucky. Most commonly, portfolio assessments are used to document writing or integrated language arts (Arter & Spandel, 1992). However, Equals (1989) and Mumme (1990) describe portfolios adopted to mathematics assessments and Collins (1991) to science assessments.

Stecher and Hamilton (1994) report the evaluation of the state-wide portfolio assessment program implemented in Vermont. They found portfolios to be unreliable for both student and school level reporting. Vermont teachers also drew attention to potential problems in the assessment implementation including the amount of time required to plan, administer and mark portfolio assessments. Stecher and Hamilton (1994) also report variation in the approach of teachers to implementing portfolios. This raises the issue of possible poor reliability, especially in a large assessment program. Arter (1989, 1991), Arter and Spandel (1992), Rothman (1990) and Valencia (1989) draw attention to other problems associated with portfolio assessments including content within the portfolio being an unrepresentative sample of the examinee's work; possible poor choice of criteria used to critique portfolio content; and the finding that conclusions drawn from the portfolio may be strongly influenced by the evaluator. Despite these serious problems portfolios have been found to have a positive effect on instruction (Stecher & Hamilton, 1994).

The Future

Regardless of whether or not the criticisms mentioned earlier in this address and directed towards traditional objective tests are true, authentic assessments also have their share of problems. Moreover, as assessments using these formats are increasingly implemented and as these formats become the focus of further research, so more problems are being revealed. For example, the structure of authentic assessment formats usually permits only a small number of questions to be administered during any single test. Therefore, it is extremely difficult to comprehensively assess a varied curriculum (and most modern curricula are varied). One solution to this problem is to develop and administer

shorter tasks thereby permitting broader curriculum coverage and improving reliability (Suen & Davey, 1990). Yet shorter tasks will also reduce the richness of the evaluative information obtained from authentic assessments. Even without reducing the length of those tasks that are currently administered, these tasks tend to exhibit only moderate reliability. For example, moderate reliability values have been reported in performance questions assessing science (Shavelson, Baxter, & Pine, 1992), performance assessments using direct writing (Breland, Camp, Jones, Morris, & Rock, 1987; Hieronymus & Hoover, 1987), and various military performance tasks (Shavelson, Mayberry, Li, & Webb, 1990). Shavelson, Baxter and Pine (1992) also report examinee performance as varying widely between tasks, thus raising concerns about the reliability of examinee marks when these marks are based on comparatively small numbers of tasks (Taylor, 1994). Another issues is the variability of test marks awarded by examiners when subjective marking procedures are used. Along with the need to increase reliability is the need for uniform testing conditions to ensure fair testing practice (Taylor, 1994). Such standardized conditions are more easily achieved with traditional testing formats. Similarly, many authentic formats require a considerable financial investment in technology which soon becomes outdated. Furthermore, authentic questions require greater resources and time to develop, administer and mark. Once more this raises the issue of ensuring adequate curriculum coverage during the assessment administration when only a few questions are used (Feinberg, 1990). Also, the desirable goal (Cole, 1988) of making large scale assessment programs time efficient, cost effective and centrally processed is more difficult for complex authentic assessments compared to simple traditional tests.

Warning, the Baby Should not be Thrown Away with the Bath Water

Education, as within any discipline, is subject to trends. There exists a disturbing tendency for whatever is in vogue within one education system to be out of vogue within another. If you speak with members and delegates who attended the Second

International Conference on Educational Evaluation and Assessment it soon becomes apparent that as objective question formats are falling from grace in the United States, so there are moves towards adopting these formats in other educational systems.

When an educational innovation is adopted on a large scale, there exists a tendency for the innovation to be adopted in place of all that has gone before. The danger inherent within this practice is that if traditional testing is completely replaced by authentic assessment programs, education stands to loose years of proven, sound and beneficial evaluation practices. Authentic assessments hold the promise of providing evaluators with a rich supply of assessment information. Yet, traditional testing practices, too, continue to offer evaluators a proven source of valuable information. Some of the criticisms leveled against traditional testing methods are valid. Yet, many practical, theoretical and technical questions remain to be answered with regard to authentic assessments -- not least of which are questions concerning reliability, which is one of the most important concerns in any evaluation program.

The continued development and adoption of the authentic formats described in this paper should not replace traditional objective question formats. Instead, authentic assessment formats should be used to augment objective formats as exemplified by the multiple-choice, true-false, matching and multiple-true-false formats, to create assessments that are more effective in evaluating those skills and abilities in which an examiner is interested. In short, the paradigms should be merged to develop powerful evaluation strategies capable of mining the rich information demonstrated by our students and overcoming the flaws present in either system of assessment. Herein lies the promise that soon we, as evaluators, may design evaluation strategies to effectively assess those hitherto unassessable higher order cognitive skills and thought processes such as reasoning, problem solving and critical thinking. For it is in these outcomes which many educators profess to be most interested (Jones & Hambleton, 1992).

Furthermore, it is these outcomes which many evaluation practitioners feel are what assessments should really be targeting.

References

Arter, J. A. (1989). Assessing communication competence in speaking and listening: A consumer's guide. Portland, OR: Northwest Regional Educational Laboratory.

Arter, J. A. (1991, April). Performance assessment: What's out there and how useful is it really? Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Arter, J. A., & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. Educational Measurement: Issues and Practice, 11(2), 36-44.

Breland, H. M., Camp. R., Jones, R. J., Morris, M. D., & Rock, D. A.(1987). Assessing writing skill (Research Monograph N. 11). New York: College Entrance Examination Board.

Bunderson, C. V., Inouye, D. K., & Olson, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 367-408). New York: Macmillan.

Cole, N. S. (1988). A realist's appraisal of the prospects for unifying instruction and assessment. In Assessment in the service of learning: Proceedings of the 1987 ETS invitational conference. Princeton, NJ: Educational Testing Service.

Collins, A. (1991). Portfolios for assessing student learning in science: A new name for a familiar idea? In A. B. Champagne, B. E. Lovitts, & B. J. Calinger (Eds.), Assessment in the service of instruction. Washington, DC: American Association for the Advancement of Science.

EQUALS. (1989). Assessment alternatives in mathematics. Berkeley, CA: Lawrence Hall of Science, University of California.

Feinberg, L. (1990). Multiple-choice and its critics: Are the alternatives any better? Commentaries from the College Board, 157, 3-15.

Gronlund, N. E. (1982). Constructing achievement tests (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hieronymus, A. N., & Hoover, H. D. (1987). Iowa Tests of Basic Skills: Writing supplement teacher's guide. Chicago: Riverside.

Hambleton, R. K., Jones, R. W., & Cadman, S. (1993). Innovations in testing and evaluation of student competencies in technical and vocational education (Laboratory of Psychometric and Evaluative Research Report No. 230). Amherst, MA: School of Education, University of Massachusetts. Report prepared for UNESCO.

Horvarth, F. G. (1991, April). Assessment in Alberta: Dimensions of authenticity. Paper presented at the meetings of the NATD/NCME, Chicago, IL.

Jones, R. W., & Hambleton, R. K. (1992). Recent advances in psychometric methods. Revista Portuguesa de Educacao, 5, 1-13.

Kellaghan, T., & Greaney, V. (1992). Using examinations to improve education: A study of fourteen African countries. World Bank Technical Paper Number 165: African Technical Department Series. Washington, DC: The World bank.

Mehrens, W. A. (1992). Performance assessment for accountability purposes. Educational Measurement: Issues and Practice, 11(1), 1-9, 20.

Melnick, D. E. (1990). Computer-based clinical situations. Evaluation and the Health Professions, 13, 104-120.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1991). The state of mathematics achievement: NAEP's 1990 assessment of the nation and the trial assessment of the states (Report No. 21-ST-04). Washington, DC: National Center for Education Statistics.

Mumme, J. (1990). Portfolio assessment in mathematics. Santa Barbara, CA: Department of Mathematics, University of California.

National Commission on Testing and Public Policy, (1990). From gatekeeper to gateway: Transforming testing in America. Chestnut Hill, MA: NCTPP, Boston College.

Rothman, R. (1990, September 12). New tests based on performance raise questions. Education Week, pp. 1, 10, 12.

Royer, J. M., Carlo, M. S., & Jones, R. W., (in review). Educational measurement in developing countries.

Semple, B. M., (1992). Performance assessment: An international experiment (Report No. 22-CAEP-06). Princeton, NJ: Educational Testing Service.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. Educational Researcher, 21(4), 22-27.

Shavelson, R. J., Mayberry, P., Li, W., & Webb, N. (1990). Generalizability of job performance measurements: Marine Corps infantryman. Military Psychology, 2, 129-144.

Stecher, B. M., & Hamilton, E. G. (1994, April). Portfolio assessment in Vermont: 1992-93: The teacher's perspective on implementation and impact. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Suen, H. K., & Davey, B. (1990, April). Potential theoretical and practical pitfalls and cautions of the performance assessment design. Paper presented at the meeting of the American Educational Research Association, Boston, MA.

Taylor, C. (1994). Assessment for measurement or standards: The peril and the promise of large-scale assessment reform. American Educational Research Journal, 31, 231-262.

Valencia, S. (1989). Assessing reading and writing: Building a more complete picture. Seattle, WA: University of Washington.

Vu, N. V., & Burrows, H. S. (1994). Use of standardized patients in clinical assessments: Recent developments and measurement findings. Educational Researcher, 23(3), 23-30.

Figure 1. Flowchart showing an Audio-Visual Context Setting question from a science teacher certification examination.

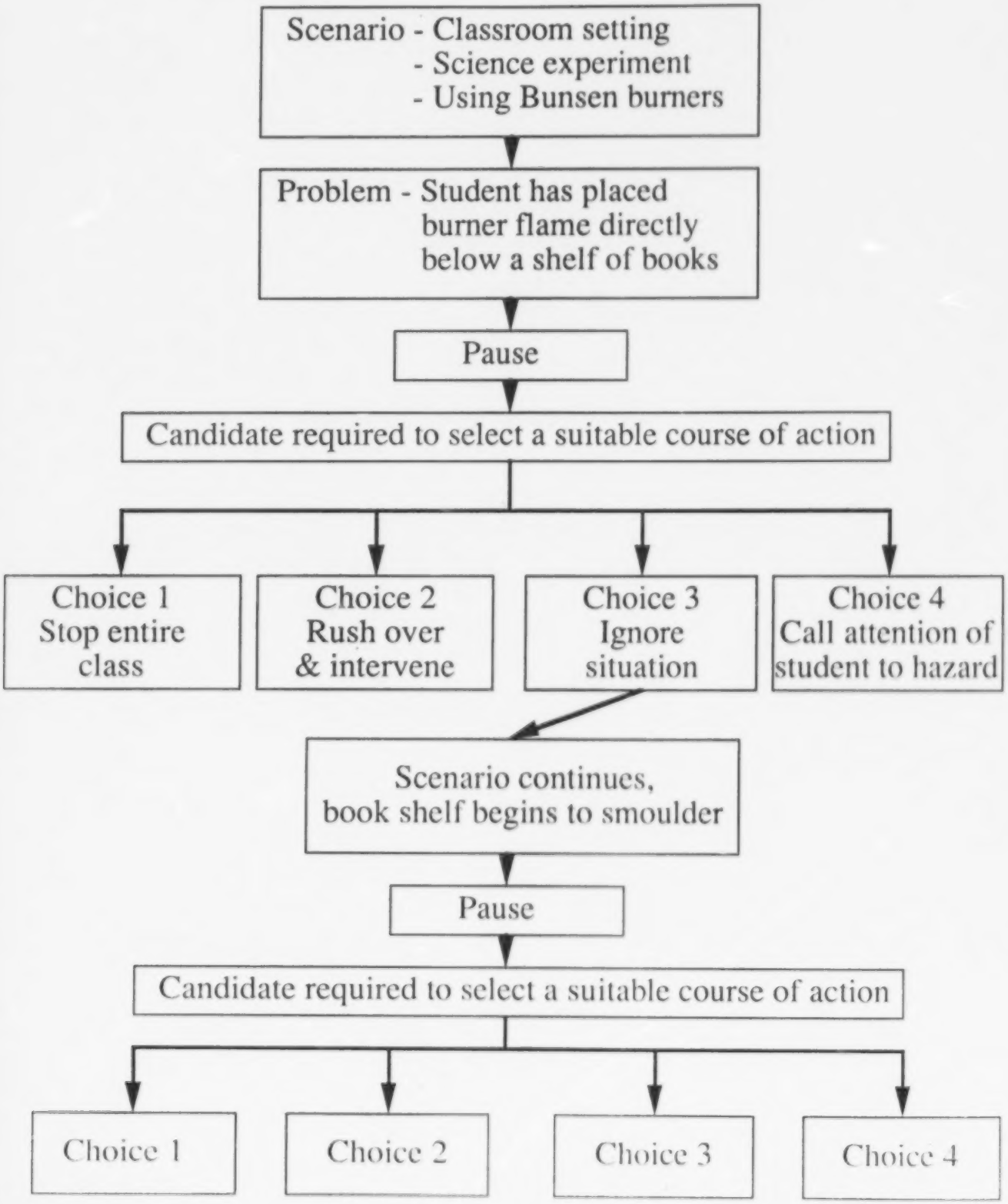
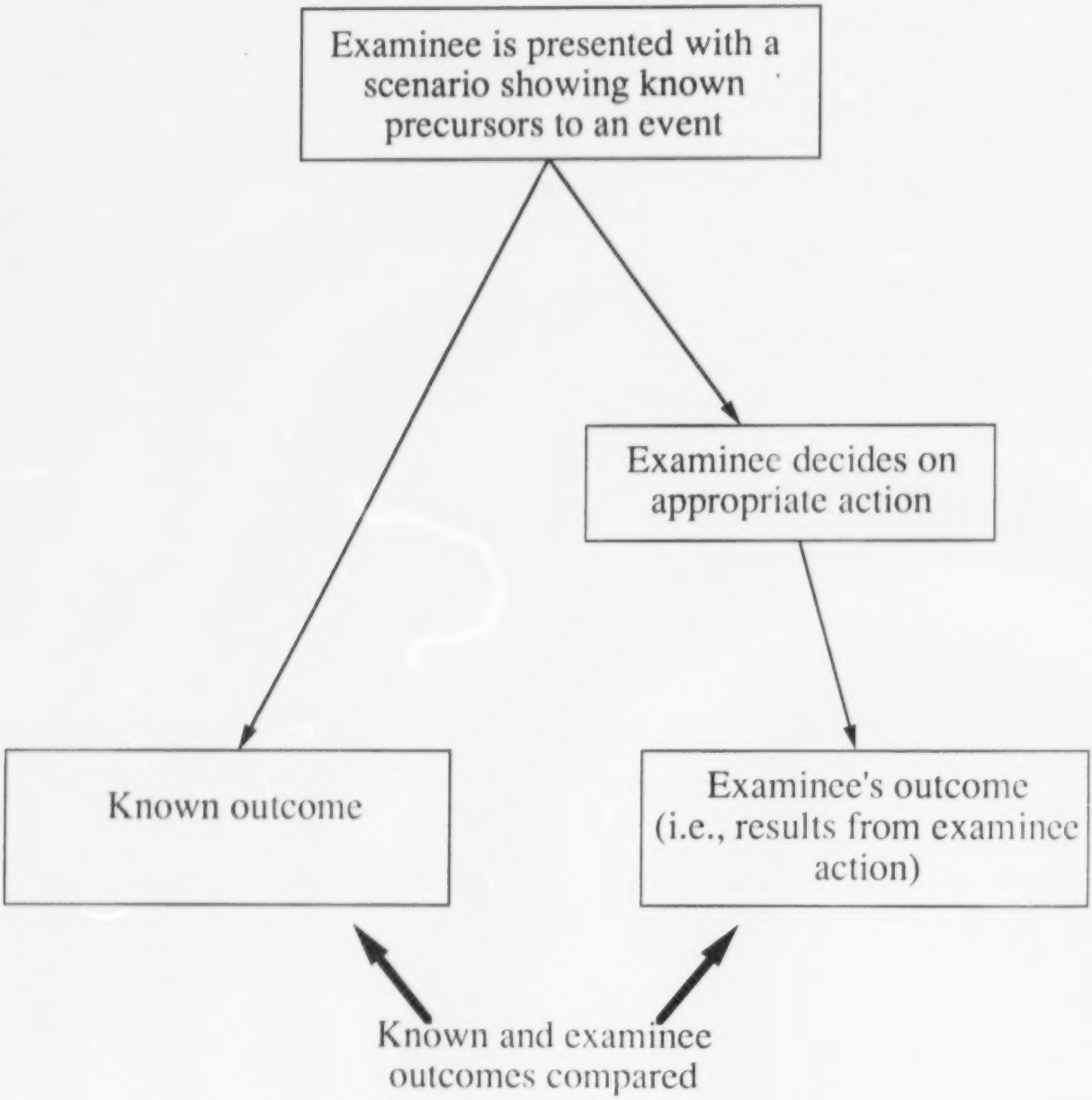

Scenario - Classroom setting
- Science experiment
- Using Bunsen burners
Problem - Student has placed burner flame directly below a shelf of books
Pause
Candidate required to select a suitable course of action
Choice 1 Stop entire class
Choice 2 Rush over & intervene
Choice 3 Ignore situation
Choice 4 Call attention of student to hazard
Scenario continues, book shelf begins to smoulder
Pause
Candidate required to select a suitable course of action
Choice 1
Choice 2
Choice 3
Choice 4

Figure 2. Flowchart showing a Computer-Based Problem Solving question.


Examinee is presented with a scenario showing known precursors to an event
Examinee decides on appropriate action
Known outcome
Examinee's outcome (i.e., results from examinee action)
Known and examinee outcomes compared